Vision Statement on Bioinformatics and Computational Chemistry

David E. Konerding, Ph.D.
Distributed Systems Department
Lawrence Berkeley National Laboratory

Overview

- Introduction
- Problems
  - Program interoperability
  - Avalanche of biodata
  - Machine reasoning on biodata
- Vision
  - Bioontology
  - Legacy application grid services
  - Bioprimitives and bioservices
- Conclusion

Introduction

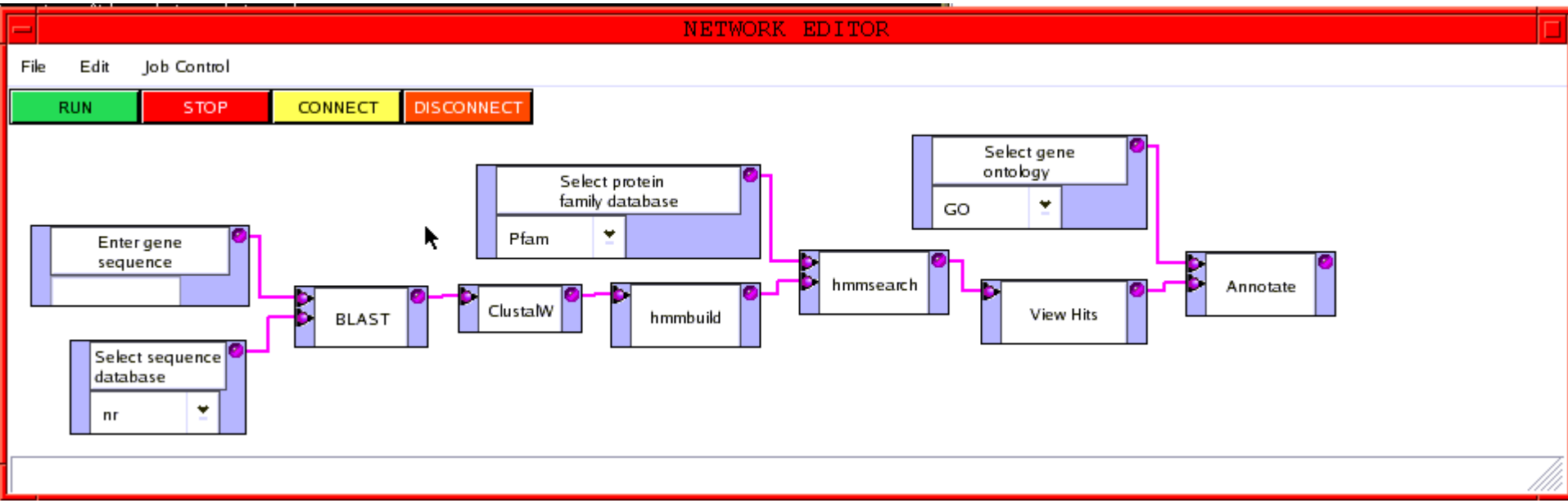
Biologists combine programs to answer questions

- Functional annotation of entire genomes
- Predict structure of all proteins in an organism
- Extract gene regulatory networks from microarray data of coexpressed genes
- Find new genes by building templates of known promoter binding sites and searching genomes

U.S. Department of Energy
Office of Science

Prototypical annotation pipeline

BERKELEY LAB



A gene annotation pipeline using BLAST to find homologs, ClustalW to align homologs, hmmer to build an HMM model and search the Protein Family Database, then annotate the results according to Gene Ontology

Problems

Legacy applications use custom formats

- Program interoperability is a big problem
  - Most interoperability focuses on data syntax, not semantics
  - Many scripts are written to automate program interoperability
    - Ad hoc
    - Not reusable
    - Brittle
- The bio* (bioperl, biopython, bioruby) toolkits address this problem but have incompatible language-specific data structures
- Biomoby (web services for bioinformatics) and S-MOBY (semantics for bioinformatics) are attempting to address these problems

Avalanche of biodata

- Ever-increasing rate of data generation:
  - Experimental methods: sequencing, structural determination, microarrays, microscopy
  - Computational: gene functional annotation and structural prediction
- Immense legacy data burden
  - Ill-defined provenance
  - nonstandard file formats
  - widely distributed across databases
- Cross-database linkage is challenging to do reliably when there is no machine-accessible metadata describing the data semantics
- Increased activity in data format standardization
  - Typically focused on data syntax rather than meaning

Machine reasoning on biodata

- Requirements for accurate machine reasoning on biodata:
  - Translation between schema
  - Data provenance
  - Service and Data discovery
  - Integration of legacy data
  - Cross-database queries
- All of these are currently handled in ad-hoc, brittle, one-off solutions.

Current approaches to biodata

- Gene Ontology (http://www.godatabase.org)
  - Successful community attempt to build an ontology of genes, gene products, and gene functions
  - Most protein database now provide links to associated GO nodes
- NCI Cancer Biomedical Informatics Grid (http://cabig.nci.nih.gov)
  - Sophisticated attempt to build controlled vocabulary and ontology to allow high-level science across multiple domains
  - Building a grid infrastructure to answer questions like:

"Annotate all the genes which are upregulated in neoplastic tissue collected from patients with increased mortality in a stomach cancer clinical trial"

Vision

Bioontology

- Different groups are developing schema covering many areas of bioinformatics and computational chemistry
  - DNA and protein sequence and structure, sequence alignments, gene structure, functional annotations, microarrays, molecular dynamics trajectories
- A common, consistent ontology will allow biologists to develop applications to ask questions at the semantic level without worrying about the data syntax

U.S. Department of Energy
Office of Science

Ontology approach to biodata

BERKELEY LAB

- No extant project constructing a consistent ontology covering the entire bioinformatics and computational chemistry domains
- Scientists need tools for ontology development
  - to model the meaning of the data
  - to visualize the consequences of applied changed ontologies to datasets
  - to assist groups developing data schema to merge their concepts

Legacy Application Grid Services

- Scientists want to ask the question “what is the function of this gene?”, not “use BLAST to search my sequence against CDD”
- Legacy applications should be exposed as web and grid services
  - Use WS-RF technology to manage resources, lifetimes, notifications
- Common queries should be exposed as bioprimitives (single applications)
- Complex queries should be exposed as bioservices (compositions of applications)
- Toolkits/frameworks to build bioprimitives and bioservices are important

Bioprimitives

- Bioprimitives expose semantic functionality of legacy applications
  - Translation of data syntax is handled under the hood-services only need to understand the semantics of the data they manage
- Bioprimitives can be used as:
  - Command line tools
  - Portlets
  - Nodes in a visual programming environment
  - Scriptable modules
- Developers should be encouraged to contribute bioprimitives to domain-specific community libraries
- Multiple bioprimitives can be composed into more complex bioservices to implement sophisticated workflows with procedural logic

U.S. Department of Energy
Office of Science

Machine reasoning using biosemantic data

BERKELEY LAB

- Once we have applied semantic meaning to biodata, we can use bioservices to perform machine reasoning
- We need to develop tools to assemble machine reasoning workflows:
  - automatic data and service discovery
  - semantic mapping of data between datasets
  - automated inference of knowledge from biodatasets using web and grid services

Conclusions

- Standardized datatypes are likely to be insufficient for integration of multiple data sources without semantics and ontologies
- Standardized ontologies are crucial for legacy bioinformatics and computational chemistry applications to interoperate reliably
- Meaning must be applied to biodata for higher-order machine reasoning to be successful.
- Collaboration tools are required to enable domain scientists to assemble vocabularies, semantics and ontologies
- Grid services are a promising approach to rapidly recompute alternative views of datasets based on differing ontologies